InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model

Year: 2024 | *Citations: 334* | *Authors: Xiao-wen Dong, Pan Zhang, Yuhang Zang*

Abstract

We introduce InternLM-XComposer2, a cutting-edge vision-language model excelling in free-form text-image composition and comprehension. This model goes beyond conventional vision-language understanding, adeptly crafting interleaved text-image content from diverse inputs like outlines, detailed textual specifications, and reference images, enabling highly customizable content creation. InternLM-XComposer2 proposes a Partial LoRA (PLoRA) approach that applies additional LoRA parameters exclusively to image tokens to preserve the integrity of pre-trained language knowledge, striking a balance between precise vision understanding and text composition with literary talent. Experimental results demonstrate the superiority of InternLM-XComposer2 based on InternLM2-7B in producing high-quality long-text multi-modal content and its exceptional vision-language understanding performance across various benchmarks, where it not only significantly outperforms existing multimodal models but also matches or even surpasses GPT-4V and Gemini Pro in certain assessments. This highlights its remarkable proficiency in the realm of multimodal understanding. The InternLM-XComposer2 model series with 7B parameters are publicly available at https://github.com/InternLM/InternLM-XComposer.